ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414349

NISP: A Multi-lingual Multi-accent Dataset for Speaker Profiling

Shareef Babu Kalluri[1], *Deepu Vijayasenan*[1], *Sriram Ganapathy*[2],
Ragesh Rajan M[1], *Prashant Krishnan*[2]

[1]National Institute of Technology Karnataka, Surathkal, India,
[2]Learning and Extraction of Acoustic Patterns (LEAP) lab, Indian Institute of Science, Bangalore.
{shareefbabu1,deepu.senan,sriram.iisc,mrageshrajan,gillyprash29}@gmail.com

Abstract

Many commercial and forensic applications of speech demand the extraction of information about the speaker characteristics, which falls into the broad category of speaker profiling. The speaker characteristics needed for profiling include physical traits of the speaker like height, age, and gender of the speaker along with the native language of the speaker. Many of the datasets available have only partial information for speaker profiling. In this paper, we attempt to overcome this limitation by developing a new dataset which has speech data from five different Indian languages along with English. The metadata information for speaker profiling applications like linguistic information, regional information, and physical characteristics of a speaker are also collected. We call this dataset as NITK-IISc Multilingual Multi-accent Speaker Profiling (NISP) dataset. The description of the dataset, potential applications, and baseline results for speaker profiling on this dataset are provided in this paper.

Index Terms: NISP dataset, Speaker profiling, Physical parameters, Voice forensics.

1. Introduction

In the recent years, speech is emerging as a reliable biometric for various commercial and surveillance applications. Speech contains the speaker identity information along with textual information, geographical information (region from where the individual belongs to) in the form of accent, age (child / teenager / adult), gender (male / female), social information, and also the emotional state of the person (angry, happy, sad, anxious etc.) [1]. Extraction of speaker related meta information is known as speaker profiling. This metadata can be used in commercial applications like voice agents and dialog systems, to deliver content targeted to the user [2]. Also, in forensic scenarios, speaker profiling could provide clues about the caller. Such applications have resulted in increased interest in area of speaker profiling [3] and it makes creation of datasets in this domain very essential. Building effective speaker profiling systems require large amount of good quality speech data along with metadata such as gender, age, physical characteristics, and accent.

Existing speech corpora has limited information about speaker metadata. Most of them have either physical characteristics or accent information, but often not about both. For example, the most common dataset TIMIT [4] has only age, height and gender information about the speakers. There is no information about other physical parameters or about the accent. The popular Speaker Recognition Evaluation (SRE) challenge datasets [5, 6, 7] have the information about smoking habits and native country. They don't have linguistic information. Other datasets such as 2010 Interspeech Paralinguistic Challenge(ComParE) dataset [2], Fisher English Corpus [8], SpeechDat II dataset [9] provide only the gender and age group information of the speaker. The CMU Kids [10] dataset only contains the grade information of the kids. None of these datasets provide any details about physical parameters beyond height and age. The only exception to this is the Copycat corpus [11] that has details of height, weight and age, but the speakers are limited to children. Similarly there are also data sets that provide the accent information of the speakers such as Accents of British Isles (ABI-1) corpus [12] and the CSLU-Foreign Accent English (FAE) [13] datasets. In this context, there is a need for dataset with richer metadata including the linguistic content for speaker profiling systems.

Another limitation of current datasets is that most of the available datasets are monolingual (English). On the other hand, multi-lingual data available (for example, the Babel dataset [14]) do not have detailed speaker profiling information.

In this paper, we describe our efforts in collecting multilingual, multi-accent dataset from five Indian states. This dataset is called NITK-IISc Multilingual Multi-accent Speaker Profiling[1] (NISP) dataset. We describe the details of the dataset in this paper along with baseline results for speaker profiling.

The rest of the paper is organized as follows. Sec. 2 describes the design and description of the dataset. Sec. 2.4 provides details about the statistics of the dataset. Sec. 3 provides the list of potential applications where NISP data can be useful. Sec. 4 contains the discussion on the baseline experimental results on physical parameter estimation. This is followed by a discussion and summary in Sec. 5.

2. Database Description

The NISP dataset creation involved collecting the speech and metadata from Indian speakers belonging to five Indian languages. The entire data collection took place over the course of a year. The speakers who participated in contributing speech data for this database consisted of students, academic staff and faculty members of different educational institutions across southern India. An informed consent is obtained from the speakers to use the data for academic and research activities.

2.1. Metadata

The linguistic, regional and physical traits are collected from each speaker along with the speech data. The metadata information collected in this dataset are the following,

[1]This dataset is publicly made available in the following address, *https://github.com/iiscleap/NISP-Dataset*. This dataset is freely available for academic and research purposes with standard license agreements.

Authorized licensed use limited to: Seikei University. Downloaded on February 23,2023 at 03:22:43 UTC from IEEE Xplore. Restrictions apply.

Table 1: *Distribution of native languages', and the number of male and female speakers in the NISP dataset*

Sl.No.	Native Language	Male	Female	Total
1.	Hindi	76	27	103
2.	Kannada	33	27	60
3.	Malayalam	35	25	60
4.	Telugu	35	22	57
5.	Tamil	40	25	65
Total Speakers		219	126	345

1. Native language (L1) of the speaker and whether the speaker can read text from L1.

2. Language used in the schooling years.

3. Second language (L2) - Most commonly spoken language other than L1.

4. Regional information: The geographic location of the native place (or the place where the subject has lived dominantly).

5. Current place of residence.

6. Physical characteristics: Age, gender of the speaker and body build parameters like height, shoulder size, and weight. The age of the speaker was noted in years and the height is measured in centimetres. The shoulder size of the speaker is measured at the widest point of shoulders between acromion bone with the individual's arms at their side in centimetres. And the weight of a speaker is measured in kilograms using standard digital weighing machine.

2.2. Speech data

The audio recordings were collected in a quiet environment like a lecture hall in each of the educational institution. All necessary precautions are taken care to avoid ambient noise, and reverberations. The speech data was collected using a high quality microphone (with Scarlett solo studio, CM25 a large diaphragm condenser microphone). The data was sampled at 44.1 kHz with a bit-rate of 16 bits per sample. In order to avoid any channel variations across recordings, all the speech samples were collected using the same microphone device.

The text data used in the reading task for the speakers were presented in the L1 language as well as in English in two different sessions. The text provided to the speakers were taken from the daily news articles as unique sentences without any contextual continuity from one sentence to another in both native language and English texts. Separately, a continuous short story section was given to the speakers in both the L1 language and English language to have contextual continuity effects in the reading task. Along with these sentences, we had also used five common sentences for every speaker. This includes two TIMIT *sa1* and *sa2* sentences and three general news article sentences in English language. Similarly two common sentences were also made in the native language text. Overall, each subject provided 20-25 unique sentences in L1 and English, 20-25 contextual sentences in L1 and English, 5 common sentences for English, and 2 sentences from L1.

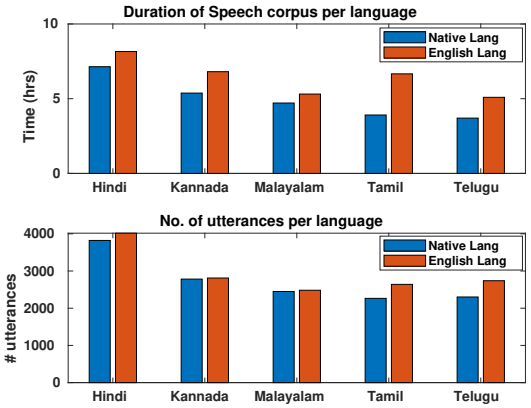

Figure 1: *Number of utterances and speech duration of each language (both native language and English speech data).*

2.3. Recording Protocol

The audio recording setup is made by using a publicly available software, namely "*Speech Recorder*" [2] and with *Focusrite Scarlett solo studio* audio recording device by connecting it to a laptop. This audio recorder device has gain controller to adjust the gain and amplitude of the speech signal while recording. The software enables a graphical user interface (GUI) to display each sentence at a time on the screen of the speaker and it is monitored and controlled by a controller on another display. The controller also verified the content, which is being read, in order to avoid any reading errors made by the speaker.

2.4. Dataset Summary

The NISP dataset has 345 speakers, which includes 219 male and 126 female speakers. The dataset has five native Indian languages (namely Hindi, Kannada, Malayalam, Tamil and Telugu) as well as Indian accented English. Each speaker provided around 4-5 minutes of speech data in each language. The distribution of speakers across the different native languages as well as gender wise distribution is shown in Table 1. The total number of utterances in this dataset is $28,268$, out of which $17,844$ are male speaker utterances, and $10,424$ are female speaker utterances. The total number of native language utterances are 13577 and there are 14691 English utterances in the dataset. This dataset has a total of 24.83 hours of native language speech data and 32.03 hours of English speech data.

The total duration of speech in hours and total number of utterances corresponding to each native language along with English speech are shown in Fig 1. The gender wise statistics of each physical parameter is given in Table 2. The total number of speakers from each region per accent is shown in Fig 2.

3. Potential Applications

The NISP dataset provides a wide range of various applications depending on the task requirement. This dataset provides the ability to explore profiling applications in text dependent or independent fashion, accent/language identification experiments, speaker recognition as well as multilingual speech recognition experiments.

[2]This software is available in this address, *https://www.bas.uni-muenchen.de/forschung/Bas/software/speechrecorder/*

Authorized licensed use limited to: Seikei University. Downloaded on February 23,2023 at 03:22:43 UTC from IEEE Xplore. Restrictions apply.

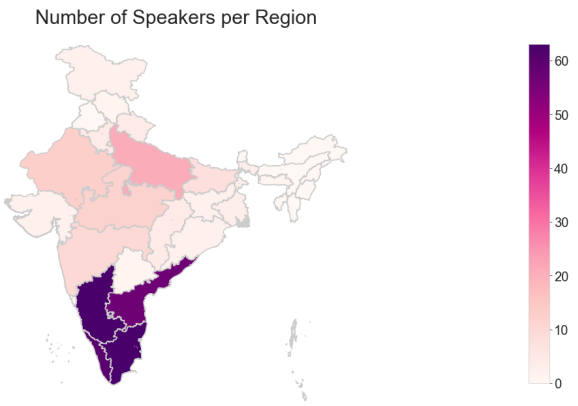

Figure 2: *Native geographic region of the speakers in the NISP dataset.*

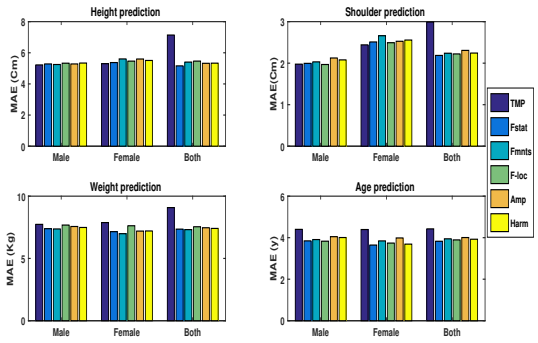

Figure 3: *Gender-wise MAE of each feature (Fstat,Formants (Fmnts), frequency locations (F-loc), Amplitude (Amp) and harmonic features (amplitude + frequency locations – Harm)) compared with Training data Mean Predictor (TMP) of the NISP dataset*

Accent & Language Identification: Identifying the accent and L1 of the speaker is an important cue in the voice forensic applications as well as in smart speaker and dialog systems. The NISP dataset enables research to explore accent related effects on speech. This database allows both L1 identification from L2 as well as language identification based on the 5 L1 languages.

Speaker Recognition: The NISP dataset, while being much smaller in scale, can be used to fine-tune the large neural network models with more multi-accent and multilingual variabilities. We hypothesize that this can improve the robustness of speaker recognition systems. In addition, multilingual speaker verification with mismatched languages in enrollment and test data can be useful for bench-marking speaker verification systems.

Speech Recognition: This dataset has potentially rich text information in both English and all the native languages (Hindi, Kannada, Malayalam, Tamil and Telugu). All these transcription, after manual verification, are recorded in UTF-8 format.

Table 2: *Gender-wise statistics of each physical parameter in the NISP dataset*

Physical Characteristic	**Min**	**Max**	**Mean**	**Standard Deviation**
Male Speakers				
Height (*cm*)	151.0	191.0	171.6	6.7
Shoulder width (*cm*)	32.0	55.0	44.7	3.2
Weight (*kg*)	43.4	116.5	69.4	11.9
Age (*y*)	18.0	47.5	24.4	5.6
Female Speakers				
Height (*cm*)	143.0	180.0	158.9	6.8
Shoulder width (*cm*)	30.0	53.0	39.7	3.4
Weight (*kg*)	34.1	86.2	56.5	10.5
Age (*y*)	18.3	46.5	25.1	6.1
Male and Female Speakers				
Height (*cm*)	143.0	191.0	166.9	9.1
Shoulder width (*cm*)	30.0	55.0	42.9	4.0
Weight (*kg*)	34.1	116.5	64.7	13.0
Age (*y*)	18.0	47.5	24.7	5.8

4. Baseline Experiments and Results

For the evaluation purposes, the dataset is divided into train and test splits without overlapping any speakers. The training split has 210 speakers with 17161 utterances, which comprises of 134 male speakers with 10911 utterances and 76 female speakers with 6250 utterances. For test split, there are 135 speakers with 11107 utterances, which includes 85 male speakers with 6933 utterances and 50 female speakers with 4174 utterances. The statistics of train and test splits of the dataset are given in Table 3. The standard error metrics like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are used to measure the errors from the actual and predicted targets.

We estimate the physical parameters like height, age, shoulder size and weight using the NISP dataset. We perform the physical parameter estimation task using three different features namely, mel filter bank features, formants and harmonics. More details about the feature extraction setup are given in [15]. We computed the first order statistics (Fstat) from the 40 Mel filter bank features using a 256 component diagonal covariance Gaussian Mixture Model Universal Background Model (GMM-UBM). The GMM was trained using 20 Mel Frequency Cepstral Coefficients (MFCC) and its deltas and double deltas together constitute 60 dimensional features. The formant and fundamental frequency features are extracted using wide band spectral components with 18^{th} order all pole model. The percentiles (5,25,50,75 and 95) are computed for the extracted features over the entire utterance. Also the harmonic features including both frequency locations (F-loc) and amplitude features (Amp.) are extracted using the narrow band spectral components using 80^{th} order all pole model. The same set of percentiles are computed for the harmonic features over the entire utterance. These computed statistics are given to linear Support Vector Regression (SVR) model to predict each physical parameter.

The MAE of each individual feature is shown in Fig 3. This is compared with the default approach - the Training data

Authorized licensed use limited to: Seikei University. Downloaded on February 23,2023 at 03:22:43 UTC from IEEE Xplore. Restrictions apply.

Table 3: *Statistics of Train and Test splits of each physical parameter in the NISP dataset*

Physical Characteristic	**Min**	**Max**	**Mean**	**Standard Deviation**
Train Speakers				
Height (*cm*)	143	191	167.1	9.5
Shoulder width (*cm*)	32	55	42.9	4.2
Weight (*kg*)	36.9	116.5	65.4	14.0
Age (*y*)	18	47.5	24.8	6.0
Test Speakers				
Height (*cm*)	146.5	182.5	166.7	8.5
Shoulder width (*cm*)	30.0	53.0	42.9	3.7
Weight (*kg*)	34.1	93.8	63.5	11.3
Age (*y*)	18.3	43.6	24.4	5.5

Mean Predictor (predicting the target physical parameter using the mean of training data of each parameter).

The three different SVR outputs of first order statistics, formants and the harmonic features (both frequency and amplitude features) were combined (Comb–3) by taking the simple average of predicted targets. This combination results in improvement of final prediction error. These results are tabulated in comparison with default predictor in Table 4. This simple average of predicted targets of these features has improved the predicted error metrics over the individual error metrics. The MAE and RMSE of both speakers (male and female speakers) improved relatively by about $22-29\%$ in body build parameter estimation (height, shoulder width and weight) tasks. Similarly, in age estimation, we observe a relative improvement of 14% improvement in MAE. There is a relative improvement over the TMP with three feature combination (Comb–3) in all the physical parameters except in RMSE of female speakers' shoulder size and male speakers' age.

We also report results on a multi-task backend framework that aims to predict the physical parameters of the dataset. The model is trained with 512 dimensional x-vector embeddings extracted using a pre-trained model. This x-vector model was trained on the VoxCeleb 1,& 2 [16, 17] corpora consisting of 7323 speakers using the extended time delay neural network (E-TDNN) architecture [18]. The x-vectors are fed to simple neural network with four feed forward layers. The first three layers are ReLU and final layer with linear activation function for predicting the physical characteristics. The model is trained on the mean square error loss. We normalise the targets while training the neural network. A separate neural networks are trained for male, female and all (both male + female) speakers. The error metrics using the x-vector model are reported in Table 4. The x-vector model is able to achieve the MAE less than the default predictor across all the physical parameters when all speakers are considered. The x-vector model showed relatively poor performance when compared with our other baseline (Comb-3). The performance degradation could be because of short utterances and smaller number of speakers to train.

5. Conclusions

A multilingual speaker profiling dataset is presented in this paper where the data was recorded in five different Indian native languages (Hindi, Kannada, Malayalam, Tamil, and Telugu) along with English language. This dataset has the linguistic information, regional information and physical characteristics of a speaker which are all useful in commercial and forensic applications of speaker profiling. This dataset has 345 (219 males and 126 females) speakers and contains $28,268$ ($17,844$ from male speaker, and $10,424$ from female speaker) utterances. Overall, this dataset has 56.86 hours of speech data in which 24.83 hours of data came from native languages of the speaker and 32.03 hours of English data. For speaker profiling tasks on this dataset, the baseline results with the combination of three features (Fstats, formants and harmonics) performs better in MAE and RMSE measures when compared to the training mean predictor.

Table 4: *Comparison of three feature combination – Comb -3 (Fstats + formant + harmonic features (amplitude + frequency locations)) with default predictor and x-vector model*

	Male	Female	All
Height (cm) Estimation			
	MAE	MAE	MAE
TMP	5.22	5.30	7.14
Comb–3	**5.16**	**5.30**	**5.11**
x-vector	5.69	6.04	5.85
Shoulder (cm) Estimation			
TMP	1.98	**2.44**	2.99
Comb–3	**1.93**	2.47	**2.11**
x-vector	2.25	3.15	2.61
Weight(kg) Estimation			
TMP	7.74	7.88	9.08
Comb–3	**7.06**	**6.84**	**7.06**
x-vector	8.37	7.56	8.03
Age(y) Estimation			
TMP	4.40	4.39	4.42
Comb–3	**3.80**	**3.55**	**3.76**
x-vector	4.01	4.94	4.39

6. Acknowledgments

This work was partially funded by Science and Engineering Research Board (SERB) under grant no: EMR/2016/007934. Authors would like to acknowledge support from institutions namely, National Institute of Technology Karnataka (NITK) Surathkal, Indian Institute of Science (IISc) Bangalore, Sree Vidyanikethan Engineering College, Tirupathi, Andhra Pradesh, KSR College of Engineering, Tiruchengode, Tamilnadu, and College of Engineering Thalassery, Kerala. We also acknowledge support from staff and students from these institutions for smooth conduction of data collection.

7. References

[1] Lawrence R Rabiner and Ronald W Schafer, *Digital processing of speech signals*, vol. 100, Prentice-hall Englewood Cliffs, NJ,

Authorized licensed use limited to: Seikei University. Downloaded on February 23,2023 at 03:22:43 UTC from IEEE Xplore. Restrictions apply.

1978.

[2] BjöRn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan, "Paralinguistics in speech and language—state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[3] Amir Hossein Poorjam, Mohamad Hasan Bahari, et al., "Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals," in *Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on*. IEEE, 2014, pp. 7–12.

[4] John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[5] "NIST Speaker Recognition Evaluation (SRE) series," https://www.nist.gov/itl/iad/mig/speaker-recognition.

[6] Alvin F Martin and Craig S Greenberg, "NIST 2008 speaker recognition evaluation: Performance across telephone and room microphone channels," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[7] Alvin F Martin and Craig S Greenberg, "The NIST 2010 speaker recognition evaluation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[8] Christopher Cieri, David Miller, and Kevin Walker, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text.," in *LREC*, 2004, vol. 4, pp. 69–71.

[9] "German SpeechDat(II)," https://catalogue.elra.info/en-us/repository/browse/ELRA-S0096.

[10] David Graff Maxine Eskenazi, Jack Mostow, "The CMU Kids Corpus," https://catalog.ldc.upenn.edu/LDC97S63.

[11] Jill Fain Lehman and Rita Singh, "Estimation of children's physical characteristics from their voices.," in *INTERSPEECH*, 2016, pp. 1417–1421.

[12] Shona M D'Arcy, Martin J Russell, Sue R Browning, and Mike J Tomlinson, "The accents of the British Isles (ABI) corpus," *Proceedings Modélisations pour l'Identification des Langues*, pp. 115–119, 2004.

[13] T Lander, "CSLU: Foreign Accented English Release 1.2," https://catalog.ldc.upenn.edu/LDC2007S08.

[14] Mary Harper, "The BABEL program and low resource speech technology," *Proc. of ASRU 2013*, 2013.

[15] Shareef Babu Kalluri, Deepu Vijayasenan, and Sriram Ganapathy, "Automatic speaker profiling from short duration speech data," *Speech Communication*, vol. 121, pp. 16–28, 2020.

[16] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[17] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[18] David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5796–5800.

Authorized licensed use limited to: Seikei University. Downloaded on February 23,2023 at 03:22:43 UTC from IEEE Xplore. Restrictions apply.